A Study of Relevance Propagation in Large Topic Ontologies

Eduardo Xamena
Knowledge Management and Information Retrieval Research Group.
LIDeCC - Laboratorio de Investigación y Desarrollo en Computación Científica (DCIC-UNS).
Planta Piloto de Ingeniería Química (UNS-CONICET) - Bahía Blanca, Argentina.
Facultad de Ciencias Exactas - UNSa - Universidad Nacional de Salta - Salta, Argentina,
Email: ex@cs.uns.edu.ar

Nélida Beatriz Brignole
LIDeCC - Laboratorio de Investigación y Desarrollo en Computación Científica (DCIC-UNS).
Planta Piloto de Ingeniería Química (UNS-CONICET) - Bahía Blanca, Argentina,
Email: nbb@cs.uns.edu.ar

Ana G. Maguitman
Knowledge Management and Information Retrieval Research Group.
LIDIA - Laboratorio de Investigación y Desarrollo en Inteligencia Artificial.
Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Bahía Blanca, Argentina,
Email: agm@cs.uns.edu.ar

Topic ontologies or Web Directories consist of large collections of links to websites, arranged by topic in different categories. The structure of these ontologies is typically not flat, since there are hierarchical and non-hierarchical relationships among topics. As a consequence, websites classified under certain topic may be relevant to other topics. While some of these relevance relations are explicit, most of them must be discovered by an analysis of the structure of these ontologies. This paper proposes a family of models of relevance propagation in topic ontologies. An efficient computational framework for inferring implicit relevance relations is described. Nine different models were computed for a portion of the Open Directory Project graph consisting of more than half a million nodes and approximately 1.5 million edges of different types. The models were initially compared from a quantitative perspective, by considering the number of inferred relations. This allowed us to distinguish the more conservative models from the less conservative ones, inducing a partial order on the set of models. A user study was carried out to compare the most promising models. It is found that some general difficulties rule out the possibility of defining flawless models of relevance propagation that only take into account structural aspects of an ontology. However, there is a clear indication that including transitive relations induced by the non-hierarchical components of the ontology results in relevance propagation models that are superior to more basic approaches.

Keywords: relevance propagation, topic ontologies, semantic similarity

Introduction

A topic ontology or Web Directory is a directory of webpages classified by topic into categories. Examples of these ontologies are Yahoo! Directory[1], Open Directory Project (ODP)[2], and their derivatives, such as Google Directory[3]. While regular Web search is the most common way adopted by users to find information on a specific topic, Web Directories are particularly useful to navigate through related topics, or when the user is not sure how to narrow her or his search from a broad category. Topic ontologies can help understand how topics within a specific area are related and may suggest terms that are useful in conducting a search. Besides being organized by topic, webpages classified in these ontologies have the advantages of having annotations (such as a description) and having been evaluated by an editor. ODP, for instance, has 20,000 volunteer editors reviewing websites and classifying them by topic.

Although Web Directories were originally conceived as a means to organize webpages to facilitate its navigation by humans, the content and structure of these directories are increasingly being used to serve other purposes. For instance, Google's regular Web search results are enhanced by information from Google Directory. ODP has been used

Corresponding Author: Ana G. Maguitman. Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur, Av. Alem 1253, 8000 Bahía Blanca, Buenos Aires, Argentina.

This research work is partially supported by CONICET (PIP 11220090100863 and 11220090100560) and Universidad Nacional del Sur (PGI 24/N029 and PGI 24/N026).

[1] http://dir.yahoo.com

[2] http://dmoz.org

[3] http://www.google.com/dirhp

to train and test automatic classifiers (Biro, Benczur, Szabo, & Maguitman, 2008; Gauch, Chandramouli, & Ranganathan, 2009), as the starting point to collect thematic material by topical crawlers (Chakrabarti, van den Berg, & Dom, 1999; Menczer, Pant, & Srinivasan, 2004), as a framework to understand the structure of content-based communities on the Web (Chakrabarti, Joshi, Punera, & Pennock, 2002), to implement information retrieval evaluation platforms (Beitzel, Jensen, Chowdhury, & Grossman, 2003; Maguitman, Cecchini, Lorenzetti, & Menczer, 2010), to understand the evolution of communities in P2P search (Akavipat, Wu, Menczer, & Maguitman, 2006), to define hierarchically-informed keyword weight propagation schemes (Kim & Candan, 2007) and to evaluate the emergent semantics of social tagging (Markines et al., 2009), among other applications. Many of these applications rely on identifying relevance or semantic similarity relationships between webpages classified in ODP.

An initial analysis of the problem of defining the relevance between documents classified in a topic ontology indicates that it essentially involves the problem of identifying non-obvious relationships from the ontology structure. Identifying these relationships in topic ontologies is a challenging problem. The structure of ontologies is typically not flat since concepts or topics can be classified according to some taxonomic schema. Topic taxonomies contain parent-child relationships between topics and their subtopics. However, relationship that go beyond the parent-child hierarchies are also common. For example, the ODP ontology is more complex than a simple tree. Some topics have multiple criteria to classify subtopics. The "Business" category, for instance, is subdivided by types of organizations (cooperatives, small businesses, major companies, etc.) as well as by areas (automotive, health care, telecom, etc.). Furthermore, ODP has various types of cross-reference links between categories, so that a node may have multiple parent nodes, and even cycles are present.

The combination of different kinds of links gives rise to intricate relations among topics. While some of these relations are explicitly given by the existing links, most of them remain implicit. Currently, ODP contains more than one million categories, making the problem of automatically deriving implicit relations between topics computationally very hard.

It is possible to define different mechanisms to derive implicit relevance relations, giving rise to multiple computational models of relevance propagation. Once relevance relations are derived, other important concepts can be defined, such as measures of similarity between topics (or between documents) in an ontology, the degree of usefulness of a document to a thematic context, or aboutness relationships between queries and topics. In particular, some widely adopted information retrieval performance measures, such as precision and recall, are defined in terms of relevance.

The goal of this article is twofold: (1) to present a family of computational models to efficiently derive implicit relevance relationships among topics from the structure of topic ontologies, and (2) to empirically evaluate these models, analyze their limitations and discuss ways to overcome them.

Background

Traditionally, the notion of relevance has been studied in the context of probability theory. In the first attempts to formalize relevance, such a notion was taken as equivalent to the notion of conditional dependence and it was subsequently refined mainly by J. M. Keynes (1921), R. Carnap (1950) and P. Gärdenfors (1978) (cited in (Gärdenfors, 1978)). A formal definition of relevance based on the use of a probability measure can be defined as follows:

Definition 1 *A formula α is relevant to a formula β given a knowledge base $\mathcal{K}$ if and only if*

$$\mathrm{P}_{\mathcal{K}}(\beta|\alpha) > \mathrm{P}_{\mathcal{K}}(\beta) \textit{ whenever } \mathrm{P}_{\mathcal{K}}(\alpha) \neq 0,$$

where $\mathrm{P}_{\mathcal{K}}$ represents a probability measure given the knowledge base $\mathcal{K}$.

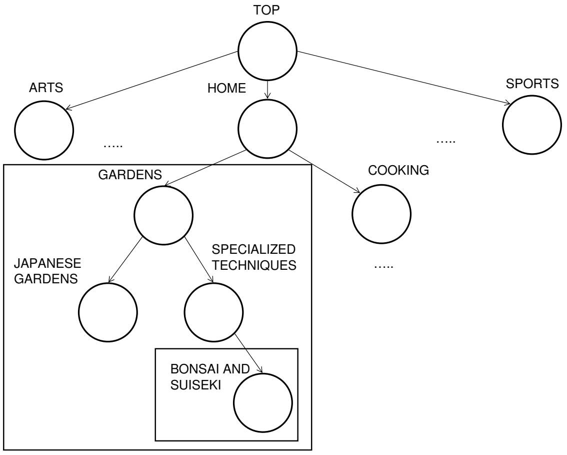


Figure 1. Illustration of a portion of a Topic Taxonomy.

In principle, adapting this definition to determine if a topic t_i is relevant to a topic t_j in a topic taxonomy $\mathcal{T}$ appears to be straightforward. The reformulation of this definition, will simply involve determining if the probability of classifying a document under topic t_j increases if we learn that the document belongs to topic t_i.

Definition 2 *A topic t_i is relevant to a topic t_j given a topic taxonomy $\mathcal{T}$ if and only if*

$$\mathrm{P}_{\mathcal{T}}(t_j|t_i) > \mathrm{P}_{\mathcal{T}}(t_j) \textit{ whenever } \mathrm{P}_{\mathcal{T}}(t_i) \neq 0,$$

where $\mathrm{P}_{\mathcal{T}}$ represents a probability measure given the topic taxonomy $\mathcal{T}$.

Given a topic taxonomy $\mathcal{T}$, we can assume $\mathrm{P}_{\mathcal{T}}(t_j)$ represents the prior probability that any document is classified under topic t_j. In practice $\mathrm{P}_{\mathcal{T}}(t_j)$ can be computed for every topic t_j in an "is-a" taxonomy by counting the fraction of documents stored in node t_j and its descendants out of all the documents in the taxonomy. The conditional probability $P_{\mathcal{T}}(t_j|t_i)$ represents the probability that any document is classified under topic t_j given that it is classified under t_i, and is computed by counting the fraction of documents stored in node t_j and its descendants out of all the documents stored in topic t_i and its descendants. In other words, $P_{\mathcal{T}}(t_j|t_i)$ is the fraction of documents in the subtree rooted at t_i that belong to the subtree rooted at t_j. For example, if the topic Bonsai_and_Suiseki is a subtopic of the topic Gardens (see figure 1), then the probability of classifying d under the topic Bonsai_and_Suiseki is higher if we know that d is classified under the more general topic Gardens than if no evidence is given in advance.

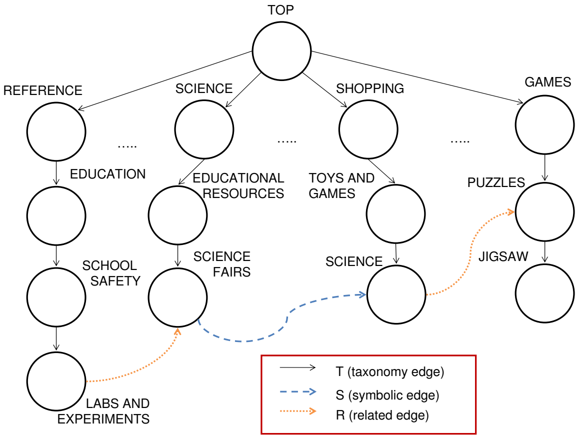


Figure 2. Illustration of a Web Directory Graph extracted from ODP.

A major limitation of definition 2 is that it is not directly applicable to general topic ontologies, such as ODP, which are more complex than a simple tree. Given a topic ontology O, the main difficulty of applying this definition remains in computing $P_O(t_j)$ and $P_O(t_j|t_i)$, as it is insufficient to count the number of documents stored in the subtrees rooted at t_i and t_j to estimate these probabilities. To illustrate this issue, take for example the topics Toys_and_Games and Puzzles in the ontology of figure 2. Although there is a clear relevance relation between these two topics, their corresponding subtrees are independent.

In a general topic ontology, computing $P_O(t_j|t_i)$ would not only involve recognizing if there is a "descendant" or "ancestor" relation between t_i and t_j. It would also involve determining if knowing that a document is related to topic t_i would have an impact on determining whether the document is about topic t_j. In other words, we need to find out in the first place whether t_i is relevant to t_j to compute $P_O(t_j|t_i)$. Therefore, for the case of a general ontology, the traditional definition of relevance relation becomes circular.

The above discussion points to the idea that defining these probabilities in terms of relevance is more natural than defining relevance in terms of probability measures. From a cognitive perspective, it is usually easier to grasp a relevance relation than to estimate probability values. Moreover, even if the probability values are given beforehand, it is possible to arrive at a wrong conclusion due to "pure numerical accidents" (R. von Mises (1963), cited in (Del Cerro & Herzig, 1996)).

To overcome the above mentioned difficulties, we will assume that *relevance is a primitive conceptual notion*. This notion, will not only capture the "is-a" relations derived from a hierarchical ontology but it will also take into consideration the non-hierarchical components. The extension of the notion of relevance from taxonomies to ontology graphs raises the question of how to extend the definition of subtree rooted at a topic for the graph case.

A "bold approach" would formulate that t_i is relevant to t_j if there is a directed path in the ontology graph from t_i to t_j. However, as we will analyze later, this formulation of topic relevance is inaccurate as the introduction of many cross links in this path can lead to a loss of meaning. In addition, allowing multiple cross links is infeasible because it leads to a dense relevance relationship, i.e., every topic becomes relevant to almost every other topic. This is also not robust because a few unreliable cross links would make significant global changes to such a relevance propagation scheme. This paper will focus on analyzing strengths and limitations of "more cautious" approaches to relevance propagation.

Related Work

Relevance is a powerful concept employed in various subdisciplines within computer science, especially in artificial intelligence and information science. This section reviews different approaches to characterize and apply relevance, and more specifically relevance propagation in the scope of knowledge management, web mining and information retrieval.

The Study of Relevance as a Key Issue in Information Science

There has been a diversity of efforts to study and characterize the notion of relevance in information science. Most research work centers on defining topical relevance, with the ultimate purpose of formulating metrics for measuring the effectiveness of information retrieval systems. An early work (Goffman, 1964) defines relevance as a measure of the information conveyed by a document relative to a query.

Although topicality has been the basis of relevance judgements in most existing proposals (as is in the present article), a number of studies have noted the inadequacy of topicality as the only ingredient in relevance judgments. For instance Rees and Saracevic (1966) argues that the definition of relevance should take into account concepts such as the information conveyed by a document, the previous knowledge of the user and the usefulness of the information to the user. Following this position, Barry (1994) highlights several user-centered criteria that affect relevance judgements. These criteria include the information content of the document, the user's previous knowledge, the user's preferences, other information and sources within the environment, the document sources, the document as a physical entity, and the user's situation. A more recent work (Xu & Chen, 2006) discusses five factors affecting relevance: topicality, novelty, reliability, understandability, and scope. After completing a user study, the authors noted that topicality and novelty are found to be the most important relevance criteria.

A more extensive review of existing literature on the concept of relevance is out of the scope of this article. The interested reader is referred to (Mizzaro, 1997), where an overview of the history of relevance in the field of information science from the 1930s to 1997 is given. More recent reviews can be found in (Saracevic, 2007b, 2007a; Hjørland, 2010).

While the notion of relevance has been addressed by several studies in information science, the notion of relevance propagation has only been partially studied. Relevance propagation becomes fundamental in the presence of interconnected structures such as subgraphs of the Web, ontologies, citation graphs and social networks in general. In particular, the notion of relevance propagation is essential for computing semantic relations between nodes arranged in any kind of network. The following sections review research work addressing these issues.

Semantic Similarity in Ontologies

Although we maintain that the notion of relevance is more primitive than the notion of semantic similarity and that the latter can be defined in terms of the former, both notions are often used interchangeably in the literature under the general name of "semantic relation". Some approaches aimed at computing measures of semantic similarity between nodes in an ontology take a network representation disregarding the taxonomical structure of the ontology. Early proposals have used path distances between the nodes in the network (e.g. (Rada, Mili, Bicknell, & Blettner, 1989)). These frameworks are based on the premise that the stronger the semantic relationship of two objects, the closer they will be in the network representation. However, as it has been discussed by several authors, issues arise when attempting to apply distance-based schemes for measuring object similarities in certain classes of networks where links may not represent uniform distances (Resnik, 1995; Jiang & Conrath, 1998; Joslyn & Bruno, 2005). In addition, some authors have argued against the suitability of relying on distance metrics when computing similarity or relevance. This is mainly due to the fact that some properties that should hold in a metric space are not valid for measures of similarity or relevance. Take for instance the triangle inequality, which is a defining property of metric space. The triangle inequality implies that if a is quite similar to b, and b is quite similar to c, then a and c cannot be very dissimilar from each other. The following example (based on William James, cited by (Tversky, 1977)) illustrates the inadequacy of this assumption: *"Jamaica is similar to Cuba (because of geographical proximity); Cuba is similar to Russia (because of their political affinity); but Jamaica and Russia are not similar at all."* This example fits the case of webpages and their topics, suggesting that the triangular inequality should not be accepted as a cornerstone of similarity or relevance models.

Another problem associated with applying distance-based approaches to compute relevance or similarity is that in hierarchical ontologies, such as ODP, certain links connect very dense and general categories while others connect more specific ones. To address this problem, some proposals estimate semantic similarity in a taxonomy based on the notion of information content (Resnik, 1995; Lin, 1998). In these approaches, the meaning shared by two objects can be measured by the amount of information needed to state the commonality of the two objects. These proposals, however, are limited to taxonomies and as a consequence do not address the question of how to estimate relevance and semantic similarity in generalized ontologies.

The general problem of computing semantic similarity in general ontologies such as the ODP graph has first been addressed in (Maguitman, Menczer, Roinestad, & Vespignani, 2005). The measure of semantic similarity proposed there takes advantage of both the ontology hierarchical ("is-a" links) and non-hierarchical (cross links) components. However, a simplistic approach to relevance propagation was taken, omitting a deep analysis of the notion of relevance and focusing only on the notion of similarity.

Computational models of semantic similarity do not need to be limited to topic ontologies and Web search. Identifying relatedness relations in other ontologies requires appropriate mechanisms to model different kinds of ontology components and their interactions. For example, the Gene Ontology[4] has two kinds of hierarchical edges ("is-a" and "part-of"). On the other hand, the WordNet ontology[5] has a much richer typology of relations. This includes semantic relations between synsets (synonym sets) such as hypernym, hyponym, meronym and holonym as well as lexical relations between senses of words (members of synsets) such as antonym, "also see", derived forms and participle.

Relevance Propagation for Identifying Topical Authoritative Sources

A variety of models of relevance propagation have been applied for identifying authoritative sources in graph representations of different domains, where graphs could represent a social network of experts, a portion of the Web, a citation network or any kind of interconnected collection of documents.

In the field of expert finding, a model of relevance propagation that relies on building an ad-hoc social network for a given query is presented in (Rode, Serdyukov, Hiemstra, & Zaragoza, 2007). The suggested framework propagates relevance through the built network to identify authorities on the required fields of expertise. A similar proposal is presented in (Serdyukov, Rode, & Hiemstra, 2008), where a graph made of both document and expert nodes is used to identify domain experts. This is accomplished by recogniz-

[4] http://www.geneontology.org/

[5] http://wordnet.princeton.edu/

ing authoritative nodes by means of a relevance propagation model.

Several approaches apply relevance propagation models to identify topic-dependant authoritative webpages, a research area known as topic distillation. For instance (Chibane & Doan, 2007) uses a traditional model of information retrieval based on content and link similarity to propagate relevance through hyperlinks. In a similar way, (Bidoki, Ghodsnia, Yazdani, & Oroumchian, 2010) proposes a content- and link-based relevance propagation model, which is iteratively enriched by information from the user's behavior. Another scheme to compute the topical authoritativeness of a webpage is presented in (Dai, Davison, & Wang, 2010). This scheme uses the ODP to build a classifier for arbitrary webpages, giving rise to a new method for authority propagation dependant on the topical relevance between the connected pages.

An alternative topic distillation method that relies on both content and link information is presented in (Shakery & Zhai, 2003) and subsequently refined in (Shakery & Zhai, 2006). In the latter work, relevance propagation through links is based on grouping neighbors into classes. A similar method is presented in (Qin et al., 2007), where instead of limiting the analysis to the hyperlinks of a web subgraph, the full structure of the sitemaps involved in the subgraph is taken into account.

Relevance Propagation in Ontologies

More closely related to our work are those frameworks that attempt to propagate relevance across topic ontologies. A model of relevance propagation in topic ontologies that takes document content into consideration is presented in (Su, Gao, Yang, & Luo, 2005). In this work, an ontology is built based on the notion of topical relevance. The resulting ontology is then used to guide a focused crawler. The ontology iteratively evolves, based on a relevance function that attempts to map the content of each discovered webpage to a class in the ontology. Relevance propagation is carried out by evolving the classes that are in the neighborhood of those classes that have been updated.

Another model of relevance propagation in topic ontologies is presented in (Kim & Candan, 2007). This work proposes a keyword propagation algorithm for augmenting the description of the entries in a navigation hierarchy by adding supplementary semantic information to the entries. In the particular case of topic taxonomies, this information is derived from the names and descriptions of the topics' ancestors and descendants. The approach is then generalized in such a way that keywords can be propagated across more complex structures.

The above two propagation schemes relate to our approach in attempting to model relevance propagation through topic ontologies. However, differently from our own framework, these proposals propagate content (e.g., keywords or keywords' weights) between pairs of neighbor entries rather than propagating relevance relations between topics across an ontology. As will be seen in the Discussion section, we contend that our proposal could be used to enhance content propagation frameworks for topic ontologies as the ones reviewed in this section.

Representing the Structure of a Web Directory Graph

A Web Directory Graph is a directed graph of nodes representing topics. Each node contains objects representing documents (webpages). A Web Directory Graph has a hierarchical (tree) component made by "is-a" links, and non-hierarchical components made by cross links of different types.

For example, the ODP ontology is a directed graph $G = (V, E)$ where:

- V is a set of nodes, representing topics containing documents;
- E is a set of edges between nodes in V, partitioned into three subsets T, S and R, such that:
- T corresponds to the hierarchical component of the ontology,
- S corresponds to the non-hierarchical component made of "symbolic" cross links,
- R corresponds to the non-hierarchical component made of "related" cross links.

Figure 2 shows a simple example of a Web Directory Graph extracted from ODP. In this graph, the set V contains topic nodes such as REFERENCE, EDUCATION, SCHOOL_SAFETY, LABS_AND_EXPERIMENTS, etc. The subset T corresponding to the hierarchical component of the Web Directory Graph contains edges such as (TOP,REFERENCE), (REFERENCE,EDUCATION), (EDUCATION,SCHOOL_SAFETY), etc. In this example there is a "symbolic" edge: (SCIENCE_FAIRS,SCIENCE) and two "related" edges: (LABS_AND_EXPERIMENTS,SCIENCE_FAIRS) and (SCIENCE,PUZZLES).

As a starting point, we say that topic t_i is relevant to topic t_j if there is an edge of some type from topic t_i to topic t_j. In the Web Directory Graph from figure 2, we can say that the topic EDUCATION is relevant to the topic SCHOOL_SAFETY, or that the topic LABS_AND_EXPERIMENTS is relevant to the topic SCIENCE_FAIRS, among other examples.

However, to derive implicit (indirect) topic relevance relations, transitive relations between edges should also be considered. An analysis of some examples leads us to conclude that while relevance relations are consistently preserved through hierarchical links, it is necessary to impose certain constraints on how the non-hierarchical links can participate in the transitive relations. Allowing an arbitrary number of cross links is infeasible because it would relate each topic to almost every other topic. Take for example the portion of ODP shown in figure 2. In this example there is a path involving three edges between topics REFERENCE/EDUCATION/SCHOOL_SAFETY/LABS_AND_EXPERIMENTS and GAMES/PUZZLES but the relevance of the first topic to the second one is questionable. On the other hand, there are other indirect paths that preserve relevance, as is the case for the path of length three between SHOPPING/TOYS_AND_GAMES and GAMES/PUZZLES/JIGSAWS.

The question addressed here is: Can we automatically derive non-obvious relevance relations among topics? Our goal is to impose certain constraints on how cross links can participate in each path in such a way that we capture the non-hierarchical components of a Web Directory Graph while preserving meaning.

In order to build our computational models of relevance propagation we start by numbering the topics in V as t_1, t_2, $\ldots$, t_n, and by representing the Web Directory Graph structure by means of adjacency matrices. Boolean matrices $\mathbf{T}$, $\mathbf{S}$ and $\mathbf{R}$ are used to codify the explicit relevance relations as described next. The matrix $\mathbf{T}$ is used to represent the hierarchical structure of an ontology. Matrix $\mathbf{T}$ codifies edges in T and is defined as $\mathbf{T}_{ij} = 1$ if $(t_i, t_j) \in T$ and $\mathbf{T}_{ij} = 0$ otherwise. The non-hierarchical components corresponding to the "symbolic" and "related" edges of the ODP graph are represented by matrices $\mathbf{S}$ and $\mathbf{R}$, respectively. Matrix $\mathbf{S}$ is defined so that $\mathbf{S}_{ij} = 1$ if $(t_i, t_j) \in S$ and $\mathbf{S}_{ij} = 0$ otherwise. The matrix $\mathbf{R}$ is defined analogously, as $\mathbf{R}_{ij} = 1$ if $(t_i, t_j) \in R$ and $\mathbf{R}_{ij} = 0$ otherwise.

Models of Relevance Propagation

Having codified the different components of the ODP graph as matrices $\mathbf{T}$, $\mathbf{S}$ and $\mathbf{R}$, we proceed to address the question of how these matrices can be used to capture the notion of relevance. Before presenting the different models of relevance propagation we review the notions of union and composition of binary relations and how these operations can be implemented as Boolean operations on matrices.

Boolean Operations on Matrices

We have already stated that relevance relations will be codified as Boolean matrices. In order to effectively compute new relations from existing ones, we have to take advantage of the existing theory that connects operations on relations with operations on matrices. In the following, we briefly review these connections.

- **Union of binary relations:** Given binary relations ρ_A and ρ_B the union $\rho_A \cup \rho_B$ can be computed as

$$\mathbf{A} \vee \mathbf{B},$$

where $\mathbf{A}$ and $\mathbf{B}$ are the matrix representations of ρ_A and ρ_B, respectively. The Boolean addition operation $\vee$ on matrices is defined as $[\mathbf{A} \vee \mathbf{B}]_{ij} = \mathbf{A}_{ij} \vee \mathbf{B}_{ij}$.

- **Composition of binary relations:** Given binary relations ρ_A and ρ_B the composition $\rho_A \circ \rho_B$ can be computed as

$$\mathbf{A} \otimes \mathbf{B},$$

where $\mathbf{A}$ and $\mathbf{B}$ are the matrix representations of ρ_A and ρ_B, respectively. The Boolean product operation $\otimes$ on matrices is defined as $[\mathbf{A} \otimes \mathbf{B}]_{ij} = \bigvee_k (\mathbf{A}_{ik} \wedge \mathbf{B}_{kj})$.

A Model Induced by Explicit Relevance Relations

Consider the logical $\vee$ operation on matrices, and let $\mathbf{M_1}$ be computed as follows:

$$\mathbf{M_1} = \mathbf{T} \vee \mathbf{S} \vee \mathbf{R} \vee \mathbf{I},$$

where $\mathbf{I}$ is the identity matrix. Matrix $\mathbf{M_1}$ is the adjacency matrix of graph G augmented with 1s on the diagonal. While matrix $\mathbf{M_1}$ accounts for all the explicit relevance relations existing in ODP it fails to capture many indirect relevance relations that result from applying transitive closures or combining relations of different types. Model $\mathbf{M_1}$ will be the most conservative of the proposed models.

Models Induced by the Transitive Closure on the Hierarchical Component

We use the Boolean product of matrices to recursively define $\mathbf{T}^{(r)}$ as follows. Let $\mathbf{T}^{(0)} = \mathbf{I}$, and let $\mathbf{T}^{(r+1)} = \mathbf{T} \otimes \mathbf{T}^{(r)}$.

Matrix $\mathbf{T}^{(r)}$ codifies all the paths of length r between topics. We define the reflexive and transitive closure of $\mathbf{T}$, denoted $\mathbf{T}^*$, as follows:

$$\mathbf{T}^* = \bigvee_{r=0}^{\infty} \mathbf{T}^{(r)}$$

Matrix $\mathbf{T}^*$ codifies all the paths (of any length) existing between pairs of topics following "is-a" links. Since there is a finite number of topics, matrix $\mathbf{T}^*$ can be computed in a finite number of steps. In this matrix, $\mathbf{T}^*_{ij} = 1$ if t_j belongs to the the topic subtree rooted at t_i, and $\mathbf{T}^*_{ij} = 0$ otherwise.

Since we have observed that relevance relations are consistently preserved through the "is-a" links it is reasonable to compute the closure $\mathbf{T}^*$ and augment it with the matrices representing the "symbolic" and "related" links. This gives rise to our second model of relevance propagation:

$$\mathbf{M_2} = \mathbf{T}^* \vee \mathbf{S} \vee \mathbf{R}.$$

In this new model, topic t_i is relevant to topic t_j if (1) there is a path from topic t_i to topic t_j involving "is-a" links only, or (2) there is a "symbolic" or "related" link from topic t_i to topic t_j. Model $\mathbf{M_2}$ is a conservative model in the sense that it propagates relevance through the hierarchical component of the ODP graph only, while the participation of cross links is restricted to explicit (direct) relevance relations.

A question that arises next is whether cross links can be included in indirect paths while preserving meaning. We have observed earlier (figure 2) that relevance is often lost if an arbitrary number of cross links are added to a path. Therefore, for the relevance propagation models to be plausible certain constraints should be imposed.

Below we formulate a family of plausible models of relevance propagation, which result from extending the previous models.

Models Induced by Propagating Cross Links throughout the Taxonomy

A simple way to incorporate cross links into the model is by propagating them upwards or downwards through the taxonomy. If we want to propagate relevance relations induced by cross links towards the root, we obtain the following model of relevance:

$$\mathbf{M_3} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I}).$$

Alternatively, if we propagate relevance relations induced by cross links towards the leaves of the taxonomy we obtain the following model:

$$\mathbf{M_4} = (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I}) \otimes \mathbf{T}^*.$$

Finally, we can propagate relevance relations induced by cross links throughout *all* the taxonomy, but allowing a single cross link in each path. This results in the following model:

$$\mathbf{M_5} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I}) \otimes \mathbf{T}^*.$$

In a previous work, model $\mathbf{M_5}$ of relevance propagation has been applied in the computation of semantic similarity measures with good results (Maguitman et al., 2005).

Another question that arises is whether relevance relations should be symmetric. The hierarchical component of the ODP graph (i.e., "is-a" links) codifies relevance relations from a parent topic to a child topic that in most of the cases are non-symmetric. In the meantime, since duplication of URLs is disallowed, "symbolic" links are a way to represent multiple memberships, for example the fact that the pages in topic SHOPPING/TOYS_AND_GAMES/SCIENCE also belong to topic SCIENCE/EDUCATIONAL_RESOURCES/SCIENCE_FAIRS. Therefore, "symbolic" links also codify parent-child relationships which, as is the case with "is-a" links, are generally non-symmetric. On the other hand, "related" links appear to codify symmetric relevance relations. Consequently, a new model of relevance can be formulated by making the "related" links bidirectional. This is achieved by extending the set of cross-link matrices with $\mathbf{R^T}$, i.e., the transpose of $\mathbf{R}$, resulting in the following model of relevance propagation:

$$\mathbf{M_6} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{R^T} \vee \mathbf{I}) \otimes \mathbf{T}^*.$$

Alternative models can be obtained by imposing additional constraints or by relaxing some. In general, "related" links appear to be weaker than the other types of links. We can reflect this in a new model that results from disallowing the downward propagation of "related" links:

$$\mathbf{M_7} = (\mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{I}) \otimes \mathbf{T}^*) \vee (\mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R^T} \vee \mathbf{I})).$$

A generalization of $\mathbf{M_7}$ is $\mathbf{M_8}$, where both "symbolic" and "related" links are allowed to simultaneously participate in the same path:

$$\mathbf{M_8} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{I}) \otimes \mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R^T} \vee \mathbf{I}).$$

There is a plethora of ways in which these models can be constrained or amplified. For example, we could allow up to n "symbolic" links as is shown in the following generalization of $\mathbf{M_8}$:

$$\mathbf{M_9} = \mathbf{T}^* \otimes (\mathbf{T} \vee \mathbf{S} \vee \mathbf{I})^n \otimes \mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R^T} \vee \mathbf{I}).$$

Figure 3 shows possible relevance paths from a source to a target node according to the different models. Various models have been considered, but the ones discussed above capture the most interesting or salient aspects of the notion of relevance propagation analyzed here.

Analyzing the Models

Quantitative Comparison

The proposed models were computed for the ODP ontology. The portion of the ODP graph we have used for our analysis consists of 571,148 topic nodes (only the WORLD and REGIONAL categories were discarded). The following table shows the size of the components of the graph used in our analysis.

Component	Size
V	571,148 nodes
T	571,147 edges
S	545,805 edges
R	380,264 edges

In order to quantitatively compare the different models, we looked at the number of relevance relations between pairs of topics induced by each model. This comparison is shown in table 1.

The above comparison table reveals a wide variation in the number of relevance relations induced by each model. In addition, we computed the number of differences among the models, and observed that for some pairs of models, such as $\mathbf{M_6}$ and $\mathbf{M_9}$, the number of differences is as large as 177,799,003.

Qualitative Analysis

Having observed that the models produced quantitatively different characterizations of the notion of relevance, we proceeded to perform an analysis of the quality of the relations induced by each.

An important theoretical observation is that the set of models form a partial order under the relation "$\mathbf{M_m} \leq \mathbf{M_n}$ if and only if $[\mathbf{M_m}]_{ij} = 1$ implies $[\mathbf{M_n}]_{ij} = 1$ for all i, j". The resulting partial order is depicted in figure 4 and can be easily shown to hold by analyzing the definition of each model as well as the definitions of the $\vee$ and $\otimes$ operators.[6]

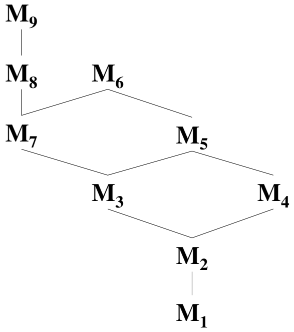


Figure 4. Partial order on the set of models.

[6] Furthermore, this is consistent with the models computed using the ODP graph.

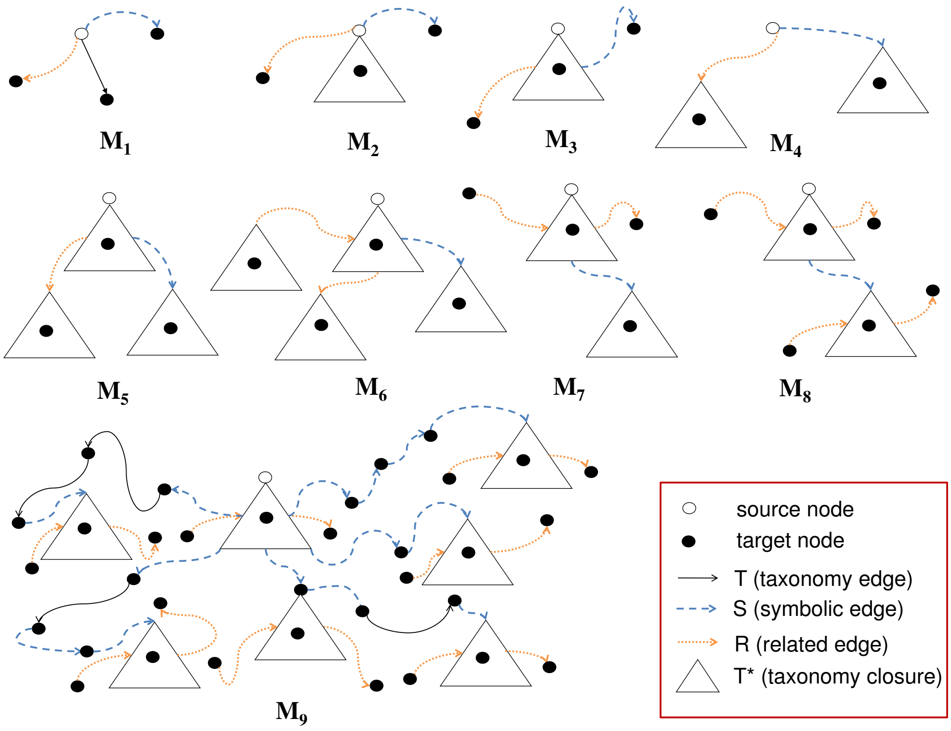


Figure 3. Possible paths from source node to target nodes in different models of relevance propagation.

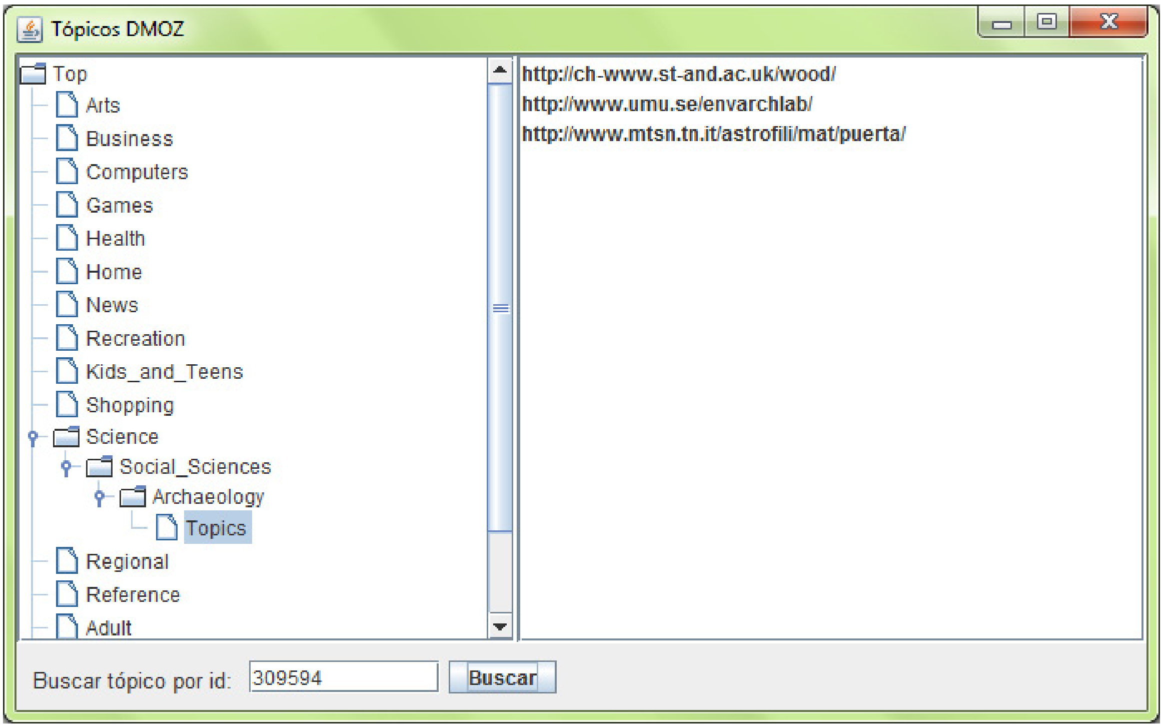
Table 1
Quantitative comparison of the models.

Model	Number of relations
$\mathbf{M_1} = \mathbf{T} \vee \mathbf{S} \vee \mathbf{R} \vee \mathbf{I}$	2,068,364
$\mathbf{M_2} = \mathbf{T}^* \vee \mathbf{S} \vee \mathbf{R}.$	5,502,581
$\mathbf{M_3} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I})$	7,072,930
$\mathbf{M_4} = (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I}) \otimes \mathbf{T}^*$	71,443,444
$\mathbf{M_5} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{I}) \otimes \mathbf{T}^*$	170,573,370
$\mathbf{M_6} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I}) \otimes \mathbf{T}^*$	174,534,253
$\mathbf{M_7} = (\mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{I}) \otimes \mathbf{T}^*) \vee (\mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I}))$	14,177,359
$\mathbf{M_8} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{I}) \otimes \mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I})$	16,915,322
$\mathbf{M_9} = \mathbf{T}^* \otimes (\mathbf{T} \vee \mathbf{S} \vee \mathbf{I})^n \otimes \mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I})$ with $n = 4$	37,609,462

In order to dig deeper into the qualitative aspects of each model, we implemented the visualization tool shown in figure 5. This tool was used in combination with the computed matrices to identify cases in which models disagreed regarding the existence or absence of a relevance relation between pairs of topics. Once conflicting topics were identified in the models, the visualization tool allowed us to visualize these topics and the set of webpages associated with them. This helped us to address the problem of which models produce the most accurate characterization of the notion of relevance.

Relevance is a highly subjective concept (Burgin, 1992; Bailey et al., 2008). After an initial pilot experiment we observed low levels of agreement in relevance judgements between the human evaluators. To further complicate the task of evaluating the different models, we noticed that even for the same judge a relevance relation that existed at a certain point of time, may disappear later, or vice versa. Despite these discrepancies, for a good number of pairs of topics there was a clear agreement concerning the existence or absence of an implicit relevance relation. For example, in figure 2 the existence of an implicit relevance relation between the topic Shopping/Toys_and_Games and the topic Games/Puzzles/Jigzaw is unquestionable, yet only models $\mathbf{M_5}$ and $\mathbf{M_6}$ capture this relation. On the other hand, there is not a clear relevance relation between the topics Society/Organizations/Students and Arts/Art_History/Movements/Impressionism in figure 6 despite the facts that the less conservative models ($\mathbf{M_5}$, $\mathbf{M_6}$,$\mathbf{M_7}$,$\mathbf{M_8}$ and $\mathbf{M_9}$), would indicate the existence of such a relation.

Instances similar to the one illustrated in figure 6 are per-



http://ch-www.st-and.ac.uk/wood/
http://www.umu.se/envarchlab/
http://www.mtsn.tn.it/astrofili/mat/puerta/

Buscar tópico por id: 309594

Figure 5. Screenshot of the Visualization Tool developed to navigate the ODP hierarchical component.

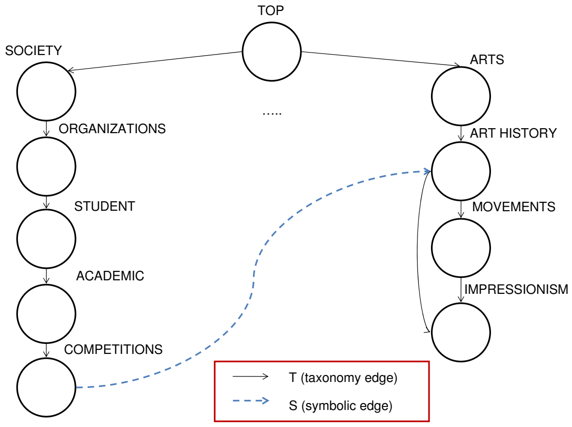


Figure 6. A questionable relation in ODP.

vasive in ODP. This highlights the fact that less conservative schemes of relevance propagation are not robust because a few unreliable cross links make significant global changes to the relevance propagation models. On the other hand, the most conservative schemes are incomplete, and hence unable to derive many useful relevance relations induced by the less conservative ones.

Validation by User Study

In order to evaluate the accuracy of some of the proposed models, we carried out an experiment to compare two of the most promising of them. The purpose of this experiment was twofold:

- In the first place, to determine if one of the analyzed models is more accurate than the other.
- Secondly, to highlight the importance of incorporating relevance relations that go beyond the basic models.

This evaluation was carried out by performing a user study that involved thirty two volunteer human subjects. Each participant was shown a sequence of thirty triplets of websites belonging to a main topic and two potentially related topics. The selection of these topics is explained later in this section. For every shown triplet, an image associated with the main topic was presented on the top of the screen and two images associated with the potentially related topics were presented below. To avoid favoring one particular model, the two images were randomly displayed one at the left and the other at the right side of the screen. Only the images of the selected websites were shown, and no information about the corresponding topics was given. The participants were given the possibility of navigating the sites. For each shown triplet, the users were asked to decide which of the candidate webpages was more related to the main page, by selecting one of the following options:

- The page on the left is more related than the page on the right to the main page.
- The page on the right is more related than the page on the left to the main page.

- Both are equally related.
- Neither is related.

The language of the websites selected for the experiment was restricted to English and therefore the participants were required to have proficiency in this language. An example of a question presented at the experiment is shown in figure 7.

Model Selection. The selection of the most promising models was conducted by considering those ones that were less conservatives, without reaching too bold models. The goal was to highlight the transitive relations between topics, avoiding too many steps that involved cross-reference links, such as is the case in M_9. Another important aspect that was considered for the selection was that most of the remaining models should be included in the selected ones (e.g, M_7 is contained in both M_6 and M_8, while M_5 is contained in M_6). The performed pilot study aided the selection process, by leading to the identification of useful relevance relations that were present in less conservative models but absent in the most basic models. Figures 8 and 9 illustrate examples of such relations. For instance M_6 induces a relevance relation between the topics Science/Physics/Instruments_And_Supplies and Science/Instruments_And_Supplies/Laboratory_Equipment/Glass_Products_And_Accesories. However, most of the proposed models are unable to identify this relation. Similarly, M_8 infers a relevance relation between the topics Business/Energy_And_Environment/Oil_And_Gas and Science/Earth_Sciences/Products_And_Services/Consulting, which is not identified by the rest of the computed models.

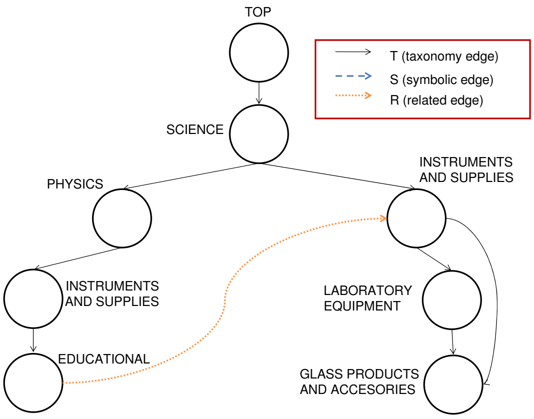


Figure 8. An example of a useful relation existing in $\mathbf{M_6}$ but absent in the other analyzed models.

Taking into account the above considerations, the selected models were:

- $\mathbf{M_6} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I}) \otimes \mathbf{T}^*$
- $\mathbf{M_8} = \mathbf{T}^* \otimes (\mathbf{S} \vee \mathbf{I}) \otimes \mathbf{T}^* \otimes (\mathbf{R} \vee \mathbf{R}^\mathbf{T} \vee \mathbf{I})$

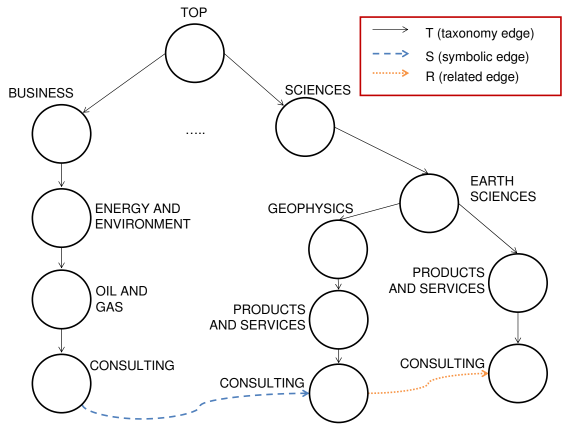


Figure 9. An example of a useful relation existing in $\mathbf{M_8}$ but absent in the other analyzed models.

Setting up the Experiment. Once $\mathbf{M_6}$ and $\mathbf{M_8}$ were selected as the most promising candidate models, the next task was to isolate topic triplets (t_1,t_2,t_3) that satisfy the following conditions:

- The main topic t_1 must have at least one related topic according to $\mathbf{M_6}$ and another related topic according to $\mathbf{M_8}$.
- The topic t_2 must be related to t_1 according to $\mathbf{M_6}$ but not according to $\mathbf{M_8}$.
- The topic t_3 must be related to t_1 according to $\mathbf{M_8}$ but not according to $\mathbf{M_6}$.

An example of a triplet satisfying these conditions can be seen in table 2.

Using matrix notation, the topic triplets (t_1,t_2,t_3) are required to satisfy the following condition:

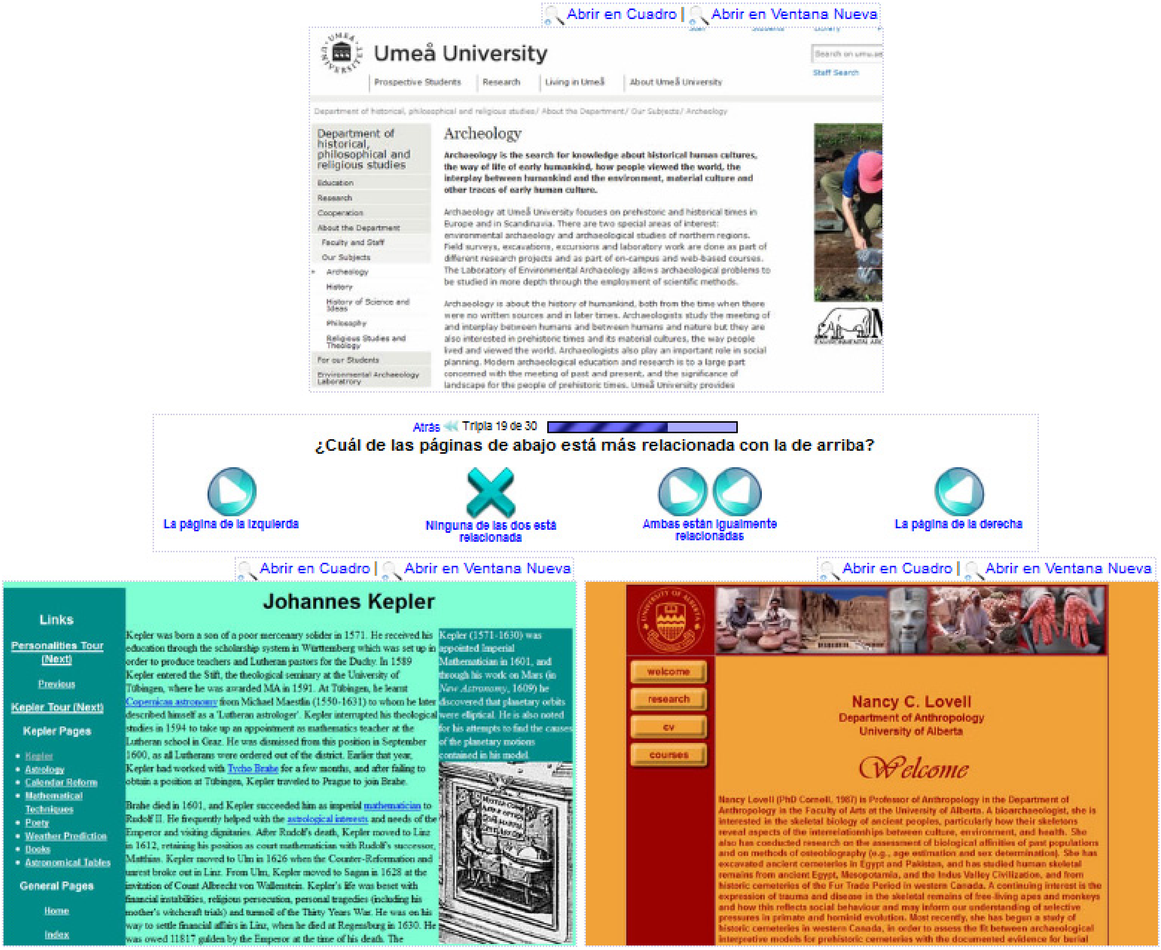
$$\mathbf{M_6}[t_1,t_2] \wedge \neg\mathbf{M_8}[t_1,t_2] \wedge \neg\mathbf{M_6}[t_1,t_3] \wedge \mathbf{M_8}[t_1,t_3]. \qquad (1)$$

The first step for identifying these triplets was to isolate those relations that were present in a model but not in the other. This was done by applying the logical minus operator on the models' matrices as follows:

- Candidate relations from $\mathbf{M_6}$: $\mathbf{M_6} \setminus \mathbf{M_8} = \mathbf{M_6} \wedge \neg\mathbf{M_8}$.
- Candidate relations from $\mathbf{M_8}$: $\mathbf{M_8} \setminus \mathbf{M_6} = \mathbf{M_8} \wedge \neg\mathbf{M_6}$.

This allowed us to identify sets of candidate relations from each model. The number of candidate relations in $\mathbf{M_6} \setminus \mathbf{M_8}$ was 159.926.121 (i.e. non-zero elements in the resulting matrix), while the number of candidate relations in $\mathbf{M_8} \setminus \mathbf{M_6}$ was 2.307.190. The resulting candidate relation matrices allowed us to isolate a sequence of triplets (t_1,t_2,t_3) satisfying condition 1. This was accomplished by searching for row indices t_1 and column indices t_2 and t_3 such that both the $(t_1,t_2)^{th}$ entry of $\mathbf{M_6} \setminus \mathbf{M_8}$ and the $(t_1,t_3)^{th}$ entry of $\mathbf{M_8} \setminus \mathbf{M_6}$ were non-zero.

With the purpose of making the experiment more accurate, we selected five main topics, each one associated with six triplets, resulting in a total of thirty triplets. In this manner, we also avoided an excessive load on the cognitive effort of human-subjects who performed the experiment, given that only after six triplets were shown they had to reassimilate the main topic subject. This methodology was similar to the one



Abrir en Cuadro | Abrir en Ventana Nueva

Umeå University

Prospective Students | Research | Living in Umeå | About Umeå University

Search on umu.se

Staff Search

Department of historical, philosophical and religious studies/ About the Department/ Our Subjects/ Archeology

Department of historical, philosophical and religious studies

Education

Research

Cooperation

About the Department

Faculty and Staff

Our Subjects

- Archeology

History

History of Science and Ideas

Philosophy

Religious Studies and Theology

For our Students

Environmental Archeology Laboratory

Archeology

Archaeology is the search for knowledge about historical human cultures, the way of life of early humankind, how people viewed the world, the interplay between humankind and the environment, material culture and other traces of early human culture.

Archaeology at Umeå University focuses on prehistoric and historical times in Europe and in Scandinavia. There are two special areas of interest: environmental archaeology and archaeological studies of northern regions. Field surveys, excavations, excursions and laboratory work are done as part of different research projects and as part of on-campus and web-based courses. The Laboratory of Environmental Archaeology allows archaeological problems to be studied in more depth through the employment of scientific methods.

Archaeology is about the history of humankind, both from the time when there were no written sources and in later times. Archaeologists study the meeting of and interplay between humans and between humans and nature but they are also interested in prehistoric times and its material cultures, the way people lived and viewed the world. Archaeologists also play an important role in social planning. Modern archaeological education and research is to a large part concerned with the meeting of past and present, and the significance of landscape for the people of prehistoric times. Umeå University provides

Atrás Tripla 19 de 30

¿Cuál de las páginas de abajo está más relacionada con la de arriba?

La página de la Izquierda | Ninguna de las dos está relacionada | Ambas están igualmente relacionadas | La página de la derecha

Abrir en Cuadro | Abrir en Ventana Nueva

Links

Personalities Tour (Next)

Previous

Kepler Tour (Next)

Kepler Pages

- Kepler
- Astrology
- Calendar Reform
- Mathematical Techniques
- Poetry
- Weather Prediction
- Books
- Astronomical Tables

General Pages

Home

Index

Johannes Kepler

Kepler was born a son of a poor mercenary solider in 1571. He received his education through the scholarship system in Württemberg which was set up in order to produce teachers and Lutheran pastors for the Duchy. In 1589 Kepler entered the Stift, the theological seminary at the University of Tübingen, where he was awarded MA in 1591. At Tübingen, he learnt Copernican astronomy from Michael Maestlin (1550-1631) to whom he later described himself as a 'Lutheran astrologer'. Kepler interrupted his theological studies in 1594 to take up an appointment as mathematics teacher at the Lutheran school in Graz. He was dismissed from this position in September 1600, as all Lutherans were ordered out of the district. Earlier that year, Kepler had worked with Tycho Brahe for a few months, and after failing to obtain a position at Tübingen, Kepler traveled to Prague to join Brahe.

Brahe died in 1601, and Kepler succeeded him as imperial mathematician to Rudolf II. He frequently helped with the astrological interests and needs of the Emperor and visiting dignitaries. After Rudolf's death, Kepler moved to Linz in 1612, retaining his position as court mathematician with Rudolf's successor, Matthias. Kepler moved to Ulm in 1626 when the Counter-Reformation and unrest broke out in Linz. From Ulm, Kepler moved to Sagan in 1628 at the invitation of Count Albrecht von Wallenstein. Kepler's life was beset with financial instabilities, religious persecution, personal tragedies (including his mother's witchcraft trials) and turmoil of the Thirty Years War. He was on his way to settle financial affairs in Linz, when he died at Regensburg in 1630. He was owed 11817 gulden by the Emperor at the time of his death. The

Kepler (1571-1630) was appointed Imperial Mathematician in 1601, and through his work on Mars (in *New Astronomy*, 1609) he discovered that planetary orbits were elliptical. He is also noted for his attempts to find the causes of the planetary motions contained in his model.

Abrir en Cuadro | Abrir en Ventana Nueva

welcome

research

cv

courses

Nancy C. Lovell

Department of Anthropology

University of Alberta

Welcome

Nancy Lovell (PhD Cornell, 1987) is Professor of Anthropology in the Department of Anthropology in the Faculty of Arts at the University of Alberta. A bioarchaeologist, she is interested in the skeletal biology of ancient peoples, particularly how their skeletons reveal aspects of the interrelationships between culture, environment, and health. She also has conducted research on the assessment of biological affinities of past populations and on methods of osteobiography (e.g., age estimation and sex determination). She has excavated ancient cemeteries in Egypt and Pakistan, and has studied human skeletal remains from ancient Egypt, Mesopotamia, and the Indus Valley Civilization, and from historic cemeteries of the Fur Trade Period in western Canada. A continuing interest is the expression of trauma and disease in the skeletal remains of free-living apes and monkeys and how this reflects social behaviour and may inform our understanding of selective pressures in primate and hominid evolution. Most recently, she has begun a study of historic cemeteries in western Canada, in order to assess the fit between archaeological interpretive models for prehistoric cemeteries with the documented evidence for burial

Figure 7. An example of a triplet shown to users on the experiment.

adopted in (Maguitman et al., 2005). Besides, each triplet was required to have active associated webpages, appropriately representing the topics' contents. Hence, the selection of such triplets was not a trivial task due to various factors, such as the disappearance of some websites during the development of the experiment. A vital piece for the triplets selection process was the visualization tool (figure 5) mentioned above, which allowed us to fast check the existence and operation of websites associated with the selected topics. Figure 7 depicts the triplet shown in table 2.

Results. The average time spent per user on performing the experiment was approximately twenty minutes. We obtained the number of answers for each of the four possible options shown with every triplet. Figure 10 shows these results grouped by user and figure 11 shows the same results grouped by triplet. The dataset used to carry out this experiment as well as the individual answers given by each user is available at `http://ir.cs.uns.edu.ar/downloads/relevance_propagation_experiment_dataset.xls`.

Table 3 shows our first analysis grouping the answers by user. From this analysis we can see that the confidence intervals for the mean number of answers associated with $\mathbf{M_6}$ and $\mathbf{M_8}$ do not overlap. At first sight, we could assume that this analysis points $\mathbf{M_6}$ as a better relevance propagation scheme. However, if we look at the overlapping of the confidence intervals for the answers associated with each of the four options, we cannot say that there is a statistically significant difference. Thus, even when the means for $\mathbf{M_6}$ and $\mathbf{M_8}$ answers are different, there is not a statistically significant difference that justifies the choice of one model over the other, due to the low significance of differences with the other answers.

If we only consider the existence or absence of a relation on each answer according to the user criterion, the results are quite different. We did this by grouping the answers that indicate the existence of some relevance relation between the main topic and any of the topics of the evaluated models, for each user. These answers are the first three options on each triplet: "The one on the left", "The one on the right", and "Both are equally related". Then, we calculated the percentage of answers that reflect the existence of a relevance relation and compared it with the percentage of answers that reflect no relation (i.e. the fourth option). This comparison is shown in table 4. The chart illustrating the total percentages of answers for each of the four options is shown in figure 12,

Table 2
Example of a triplet used in the evaluation

Model	URL	Topic

Both	`http://www.idesam.umu.se/english/about/`
	`subjects/archeology/?languageId=1`
		Science
		Social_Sciences
		Archaeology
		Topics

M_6	`http://www.hps.cam.ac.uk/starry/kepler.html`	Top Science Astronomy History People Kepler,_Johannes
M_8	`http://www.ualberta.ca/~nlovell/index.html`	Top Science Social_Sciences Archaeology Archaeologists Bioarchaeologists

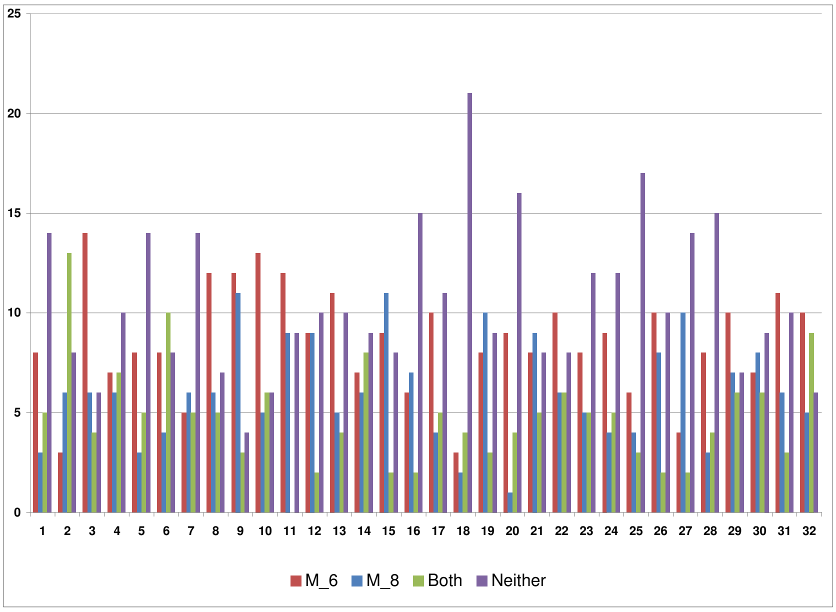


Figure 10. Number of answers for each option per user.

while the chart with the totals for the grouped options according to the second analysis is shown in figure 13.

These results indicate that there is a statistically significant difference between the means of the two groups, given that the confidence intervals do not overlap, with a significance level of 5% (95% of confidence level). As a consequence, we have enough statistical evidence to conclude that the relevance relations determined by the evaluated models are consistent in many cases according to the users' criterion, and can be taken into account for the computation of semantic similarity between websites. In other words, the basic models are insufficient to reflect useful relevance relations that could be contributed by some of the less conservative models.

Discussion

The above analysis leads us to conclude that while some models are better predictors than others of the existence or

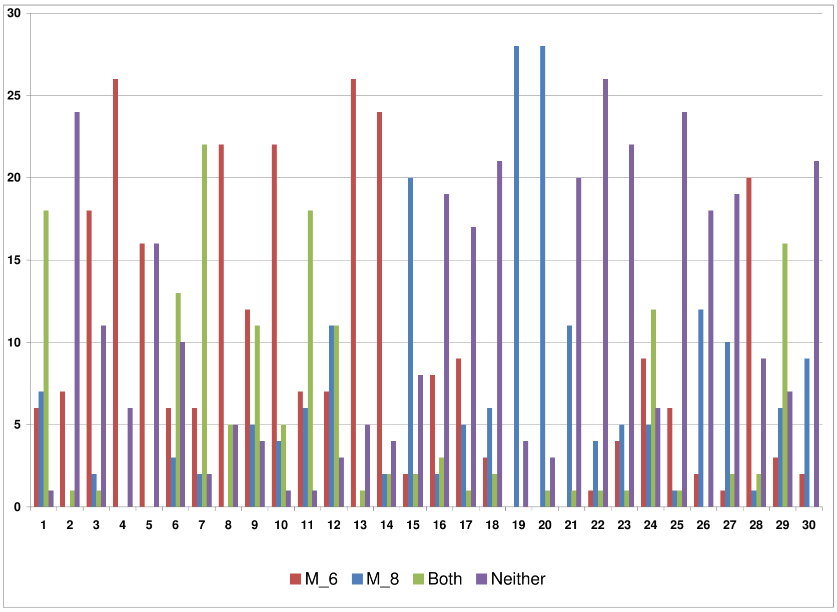


Figure 11. Number of answers for each option per triplet.

Table 3
First analysis of the experiment data.

ANSWER	N	MEAN	STDEV	95% CI
$\mathbf{M_6}$	32	28.65%	8.99%	(25.53%,31.76%)
$\mathbf{M_8}$	32	20.31%	8.61%	(17.33%,23.29%)
both	32	15.94%	8.71%	(12.92%,18.95%)
neither	32	35.10%	12.64%	(30.72%,39,48%)

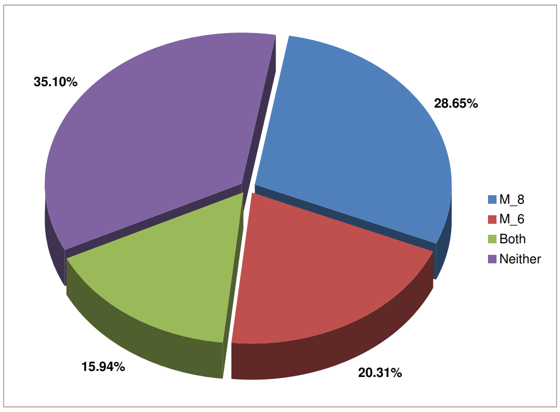


Figure 12. Percentage of answers for each option.

absence of relevance relations, none of them is flawless. This points to the fact that despite being a key concept in artificial intelligence and information science, relevance is a fuzzy and subtle notion, difficult, if not impossible, to formalize using structural aspects only.

Despite these limitations, our analysis indicates that there is a clear increase in the amount of useful information inferred when the less conservative models (such as $\mathbf{M_6}$ or $\mathbf{M_8}$) are used to identify implicit relevance relations. This analysis provides new insight into the problem of computing semantic similarity measures for general ontologies, highlighting the benefits of taking advantage of both the hierarchical and non-hierarchical components of these ontologies.

As it has been proposed in (Maguitman et al., 2005), the semantic similarity between two topics t_i and t_j in an ontology graph can be computed using an information-theoretic approach as follows:

$$\sigma^G(t_i, t_j) = \max_k \frac{2 \cdot (\mathbf{M}[t_k, t_i] \wedge \mathbf{M}[t_k, t_j]) \cdot \log \mathrm{P}_O(t_k)}{\log(\mathrm{P}_O(t_i|t_k) \cdot \mathrm{P}_O(t_k)) + \log(\mathrm{P}_O(t_j|t_k) \cdot \mathrm{P}_O(t_k))}.$$

Table 4
Second analysis of the experiment data.

ANSWER	N	MEAN	STDEV	95% CI
related ($\mathbf{M_6}$, $\mathbf{M_8}$ or both)	32	64.90%	12.64%	(60.52%,69.28%)
not related (neither)	32	35.10%	12.64%	(30.72%,39,48%)

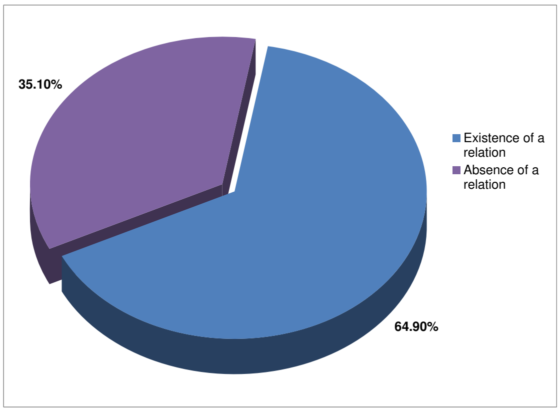


Figure 13. Percentage of answers comparing the existence of a relevance relation (coming from $\mathbf{M_6}$, $\mathbf{M_8}$ or both) with the non existence of a relevance relation.

The probability $\mathrm{P}_O(t_k)$ represents the prior probability that any document is classified under topic t_k. Once a model of relevance propagation $\mathbf{M}$ has been computed, $\mathrm{P}_O(t_k)$ can be naturally estimated in terms of model $\mathbf{M}$ as:

$$\mathrm{P}_O(t_k) = \frac{\sum_{t_j \in V}(\mathbf{M}[t_k, t_j] \cdot |t_j|)}{|U|}, \quad (2)$$

where $|t_j|$ is the number of documents directly associated with topic t_j and $|U|$ is the total number of documents in the ontology. The conditional probability $\mathrm{P}_O(t_i|t_k)$ represents the probability that any document will be classified under topic t_i given that it is classified under t_k, and it can also be estimated in terms of model $\mathbf{M}$ as follows:

$$\mathrm{P}_O(t_i|t_k) = \frac{\sum_{t_j \in V}((\mathbf{M}[t_i, t_j] \wedge \mathbf{M}[t_k, t_j]) \cdot |t_j|)}{\sum_{t_j \in V}(\mathbf{M}[t_k, t_j] \cdot |t_j|)}. \quad (3)$$

Equations 2 and 3 are in accordance with the arguments presented in the Background section, where we claim that *relevance is a primitive conceptual notion* and suggest that defining $\mathrm{P}_O(t_j)$ and $\mathrm{P}_O(t_j|t_i)$ in terms of relevance is more natural than defining relevance in terms of these probability measures.

There are a number of ways in which the proposed models of relevance propagation can be improved. For instance, the less conservative models could be combined with mechanisms that prevent them from deriving relevance relations between two topics unless an analysis of the topics' content suggests a connection between them. This analysis could be based on the text describing the topics, which is available in ODP. Another source of content are the features of the websites associated with the topics, such as the text, the outgoing links, the incoming links or a combination of all.

Another possible improvement is the extension of the proposed models to fuzzy models of relevance propagation. Different types of edges have different roles, and one way to distinguish these roles is to assign them weights. Then, the weight $w_{ij} \in [0, 1]$ for an edge between topic t_i and t_j can be interpreted as an explicit measure of the degree of membership of t_j in the family of topics rooted at t_i. In order to propagate relevance, the Boolean product of matrices $\otimes$ will need to be replaced by some fuzzy operator. For example, we could use the MaxProduct fuzzy composition operator (Kandel, 1986) defined on matrices as follows:

$$[\mathbf{A} \odot \mathbf{B}]_{ij} = \max_k(\mathbf{A}_{ik} \cdot \mathbf{B}_{kj}).$$

The element $\mathbf{M}[t_i, t_j]$ resulting from propagating relevance in the new fuzzy models will be interpreted as a fuzzy relevance relation of topic t_i to topic t_j. For certain weighting schemes, the distance between two topics in the directory will have an impact on their relevance value.

Finally, it is important to distinguish the propagation of relevance relations from the propagation of keywords (and keywords' weights) through a topical structure. Two approaches for keyword propagation (Su et al., 2005; Kim & Candan, 2007) were reviewed in the Related Work section. In these approaches, keywords are propagated through topics following the hierarchical component of a topic directory or to neighbor topics. We contend that the propagation mechanism could be extended guided by our models of relevance propagation. In other words, more complex propagation schemes can be implemented if content is propagated from topic t_i to topic t_j whenever $\mathbf{M}[t_i, t_j] \neq 0$ for a given model $\mathbf{M}$.

Conclusions

This paper addressed the problem of inferring relevance relations between topics in a Web Directory Graph by looking at structural features of the graph only. We proposed nine different models of relevance propagation and computed them for a huge graph consisting of more than half a million nodes. This resulted in a challenging computational task, for which we implemented dedicated efficient algorithms. The resulting models were compared from both a quantitative and qualitative perspective. In addition, a user study was carried out to compare two of the most promising models.

While some models appear to better approximate the notion of relevance than others, certain general difficulties appear to rule out the possibility of defining precise models of relevance propagation by considering structural aspects only. This result has interesting practical and theoretical consequences as many existing methods attempt to identify implicit semantic relations in network representations by looking only at the structure or topology of the network (e.g., (Pedersen, Patwardhan, & Michelizzi, 2004; Rada et al., 1989)). This calls for the investigation and development of mechanisms that integrate structural aspects with other aspects (such as content or other contextual aspects) to derive enhanced models of relevance propagation.

In this sense, structure and content analysis can be usefully integrated in two ways. Firstly, the proposed structural models of relevance propagation can be enhanced by taking content into consideration. Secondly, existing models of content propagation such as the ones proposed in (Su et al., 2005; Kim & Candan, 2007) (discussed in the Related Work section) can be reformulated to propagate keywords and their weights through new paths induced by the models of relevance propagation.

To the authors' knowledge, this is the first attempt to model the problem propagating relevance relations in a Web Directory Graph. The applicability of the proposed models of relevance propagation to the area of artificial intelligence and information science is extensive and multifarious. Since much of a reasoner's knowledge can be expressed in terms of relevance relations, a computational model of relevance propagation is a useful tool for the design of common-sense reasoning and information seeking systems.

References

Akavipat, R., Wu, L.-S., Menczer, F., & Maguitman, A. G. (2006). Emerging semantic communities in peer web search. In *Proceedings of the International Workshop on Information Retrieval in Peer-to-peer Networks* (pp. 1–8). New York, NY, USA: ACM.

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., Vries, A. P., & Yilmaz, E. (2008). Relevance assessment: Are judges exchangeable and does it matter. In *Proceedings of the 31st Annual International acm sigir Conference on Research and Development in Information Retrieval* (pp. 667–674). New York, NY, USA: ACM.

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, *45*, 149–159.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., & Grossman, D. (2003). Using titles and category names from editor-driven taxonomies for automatic evaluation. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management* (pp. 17–23). New York, NY, USA: ACM.

Bidoki, A. M. Z., Ghodsnia, P., Yazdani, N., & Oroumchian, F. (2010). A3crank: An adaptive ranking method based on connectivity, content and click-through data. *Information Processing and Management*, *46*(2), 159–169.

Biro, I., Benczur, A., Szabo, J., & Maguitman, A. (2008). A comparative analysis of latent variable models for web page classification. In *Proceedings of the 2008 Latin American Web Conference* (pp. 23–28). Washington, DC, USA: IEEE Computer Society.

Burgin, R. (1992, July). Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing and Management*, *28*, 619–627.

Chakrabarti, S., Joshi, M. M., Punera, K., & Pennock, D. M. (2002). The structure of broad topics on the Web. In *Proceedings of the 11th International Conference on World Wide Web* (pp. 251–262). New York, NY, USA: ACM.

Chakrabarti, S., van den Berg, M., & Dom, B. E. (1999). Focused crawling: A new approach to topic-specific web resource discovery. *Computer Networks*, *31*, 1623–1640.

Chibane, I., & Doan, B.-L. (2007). Relevance propagation model for large hypertext document collections. In *Large scale semantic access to content (text, image, video, and sound)* (pp. 585–595). Paris, France: Le centre de hautes etudes internationales d'informatique documentaire.

Dai, N., Davison, B., & Wang, Y. (2010). Mining neighbors' topicality to better control authority flow. In C. Gurrin et al. (Eds.), *Ecir* (Vol. 5993, p. 653-657). Springer.

Del Cerro, L. F., & Herzig, A. (1996). Belief change and dependence. In *Proceedings of the 6th Conference on Theoretical Aspects of Rationality and Knowledge* (pp. 147–161). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Gärdenfors, P. (1978). On the logic of relevance. *Syntheses*, *37*(3), 351–367.

Gauch, S., Chandramouli, A., & Ranganathan, S. (2009, January). Training a hierarchical classifier using inter document relationships. *Journal of the American Society for Information Science and Technology*, *60*, 47–58.

Goffman, W. (1964). On relevance as a measure. *Information Storage and Retrieval*, 201-203.

Hjørland, B. (2010). The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technologycec*, *61*(2), 217-237.

Jiang, J. J., & Conrath, D. W. (1998). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the international conference on research in computational linguistics (rocling x).* Taiwan.

Joslyn, C., & Bruno, W. J. (2005). Weighted pseudo-distances for categorization in semantic hierarchies. In *International conference on conceptual structures. lecture notes in computer science 3956* (pp. 381–395).

Kandel, A. (1986). *Fuzzy mathematical techniques with applications.* Boston, MA, USA: Addison-Wesley.

Kim, J. W., & Candan, K. S. (2007). Leveraging structural knowledge for hierarchically-informed keyword weight propagation in the web. In *Proceedings of the 8th knowledge discovery on the web international conference on advances in web mining and web usage analysis* (pp. 72–91). Berlin, Heidelberg: Springer-Verlag.

Lin, D. (1998). An Information-theoretic Definition of Similarity. In *Proceedings of the fifteenth international conference on machine learning* (pp. 296–304). San Franciso, CA, USA: Morgan Kaufmann Publishers Inc.

Maguitman, A. G., Cecchini, R. L., Lorenzetti, C. M., & Menczer, F. (2010, October). Using topic ontologies and semantic similarity data to evaluate topical search. In C. von Lücken, M. E. García, & C. Cappo (Eds.), *Xxxvi conferencia latinoamericana de informática.* Asunción, Paraguay: Facultad

Politécnica – Universidad Nacional de Asunción and Universidad Autónoma de Asunción.

Maguitman, A. G., Menczer, F., Roinestad, H., & Vespignani, A. (2005). Algorithmic detection of semantic similarity. In *Proceedings of the 14th international conference on world wide web* (pp. 107–116). New York, NY, USA: ACM.

Markines, B., Cattuto, C., Menczer, F., Benz, D., Hotho, A., & Stumme, G. (2009). Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on world wide web* (pp. 641–650). New York, NY, USA: ACM.

Menczer, F., Pant, G., & Srinivasan, P. (2004, November). Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology (TOIT)*, *4*(4), 378–419.

Mizzaro, S. (1997, September). Relevance: the whole history. *Journal of the American Society for Information Science*, *48*(9), 810–832.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). Wordnet::similarity: Measuring the relatedness of concepts. In *Demonstration papers at hlt-naacl 2004* (pp. 38–41). Stroudsburg, PA, USA: Association for Computational Linguistics.

Qin, T., Liu, T.-Y., Zhang, X.-D., Feng, G., Wang, D.-S., & Ma, W.-Y. (2007). Topic distillation via sub-site retrieval. *Information Processing and Management*, *43*(2), 445–460.

Rada, R. R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*, *19*(1), 17–30.

Rees, A., & Saracevic, T. (1966). *The measurability of relevance*. Center for Documentation & Communication Research, Western Reserve University.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Ijcai'95: Proceedings of the 14th international joint conference on artificial intelligence* (pp. 448–453). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Rode, H., Serdyukov, P., Hiemstra, D., & Zaragoza, H. (2007). *Entity ranking on graphs: Studies on expert finding.*

Saracevic, T. (2007a, November). Relevance: A review of the literature and a framework for thinking on the notion in information science (part iii: Behavior and effects of relevance). *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144.

Saracevic, T. (2007b, November). Relevance: A review of the literature and a framework for thinking on the notion in information science (part ii: Nature and manifestations of relevance). *Journal of the American Society for Information Science and Technology*, *58*(13), 1915–1933.

Serdyukov, P., Rode, H., & Hiemstra, D. (2008). Modeling multistep relevance propagation for expert finding. In J. G. Shanahan et al. (Eds.), *Cikm* (pp. 1133–1142). ACM.

Shakery, A., & Zhai, C. (2003). Relevance propagation for topic distillation UIUC TREC 2003 web track experiments. In *Trec* (pp. 673–677).

Shakery, A., & Zhai, C. (2006). A probabilistic relevance propagation model for hypertext retrieval. In *Proceedings of the 15th acm international conference on information and knowledge management* (pp. 550–558). New York, NY, USA: ACM.

Su, C., Gao, Y., Yang, J., & Luo, B. (2005). An efficient adaptive focused crawler based on ontology learning. In N. Nedjah, L. de Macedo Mourelle, A. Abraham, & M. Köppen (Eds.), *His* (pp. 73–78). IEEE Computer Society.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Xu, Y. C., & Chen, Z. (2006, May). Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology*, *57*(7), 961–973.